Biopathways and Protein Interaction Databases

Frank Olken

Lawrence Berkeley National Laboratory

PGA Course at LBNL

Friday, Feb. 28, 2003

Protein Interaction Databases

- Protein interactions from experiments
- Used to help construction biopathways DB
- Yeast 2 Hybrid Experiments:
  - pairwise interactions (i.e., graphs)
- Mass spectrometry experiments:
  - identify protein complexes
  - sets of interacting proteins (i.e., hypergraphs)
  - NOT pairwise interactions
- PIN = Protein Interaction Network

Biopathways Databases

- Metabolic pathways
- Signaling pathways
- Gene regulatory networks
- Inferred from:
  - protein interaction networks
  - micro-array data (gene expression data)
  - proximity on genome
  - other experiments
- BP = Biopathways

Outline of Talk

- Examples of BP and PIN Databases
- DB contents
- DB Uses
- Graph data model
- Graph queries

Biopathways Database Contents

- **Networks of chemical reactions**
- **Metabolic pathways: bulk reactions, ODE's**
- **Signaling pathways: rarer reactants, stochastic Petri nets**
- **Gene regulatory networks**
  - **gene expression**
- **Protein Interaction networks**
  - **protein interactions**
  - **data is very noisy**

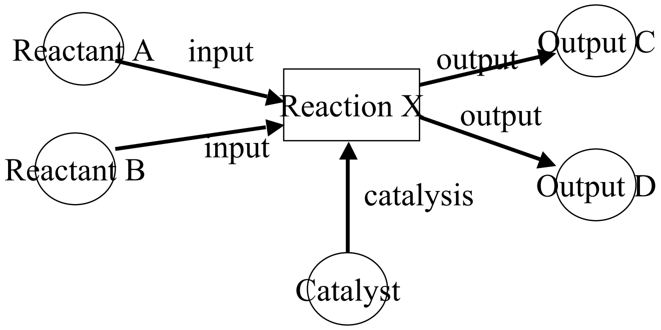
What good are biopathways databases?

- Formal encoding of metabolic, signaling, gene regulatory networks
- Framework to hang more detailed kinetics data, mathematical models
- Basis for linear programming models of cell metabolism and growth

What good are biopathways databases? (cont.)

- Comparative analysis of pathways
  - Pathogenicity studies
  - Bioengineering of organisms
  - Phylogeny
- Drug target identification
- Dietary therapy identification
- Drug metabolite identification -
  - possible adverse drug reactions

Bipartite Graph representation of chemical reaction



Directed vs. Undirected Graphs

- Biopathways = networks of chemical reactions
  - **also gene expression …**
  - **directed graphs**
  - **nodes = reactions (bioprocesses) or chemical entities**
  - **edges = relationships (inputs, outputs, catalyst …)**
- Protein interaction networks
  - **undirected graphs (for Y2H data)**
  - **nodes = proteins**
  - **undirected edges connect interacting proteins**

Pathway vs. Path

- Pathway:
  - typically a small connected subgraph of a larger graph (e.g., the entire metabolism of a microbe)
  - biological term
- Path:
  - a connected linear graph, i.e., no branches or cycles
  - A---->B---->C----->D
  - a term from graph theory

Graph models

- Simple graph = (nodes, edges)
  - edges connect a pair of nodes
- Nested graph = Russian dolls
  - graphs inside nodes
  - used for hierarchical pathways representation
  - used for chemical structure graphs for compounds

More graph models

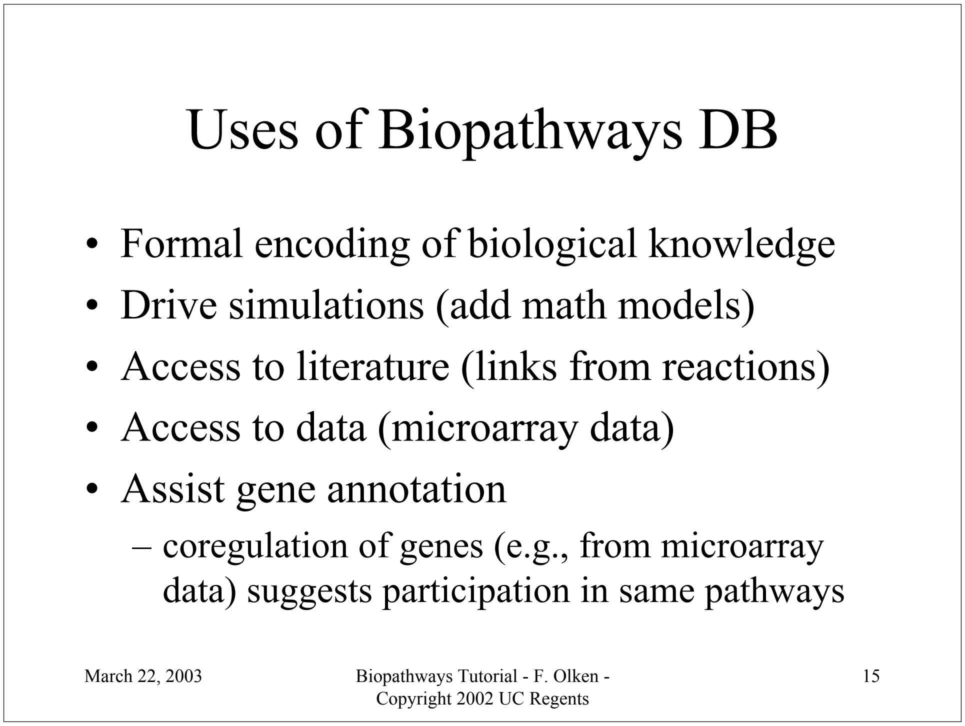
- Hypergraphs
  - edges connect sets of nodes
  - reaction = edge connecting set of inputs to set of outputs
  - more concise, more difficult for algorithms
  - used by Ozsoyoglu's at Case Western

Details of graph models

- Attributes
  - of nodes, edges (?)
  - attributes representation as:
    - tuples or structures at nodes
    - node ----attribute edge----> attribute value
- Types on nodes/edges

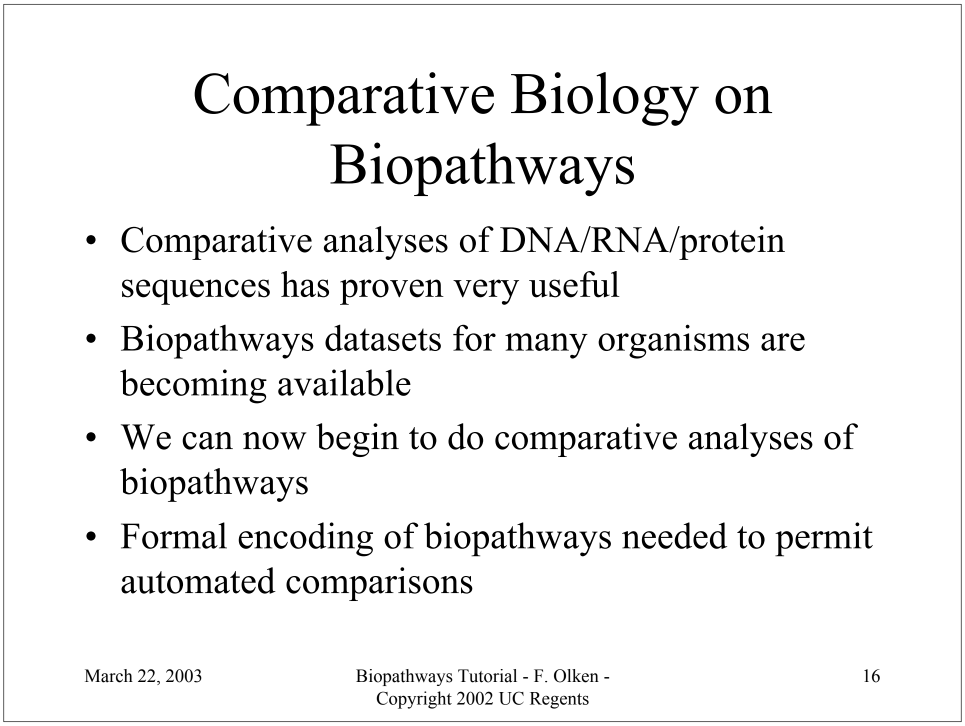
Details of graph models

- Can edges be referenced ?
- Bipartite Graph Alternative:
  - replace edges (----->) with pair of edges and a node (----> node ---->)
  - node carries attributes, can be pointed to
  - Now have two types of nodes:
    - original nodes
    - “edge nodes”
    - Edges always connect different types of nodes



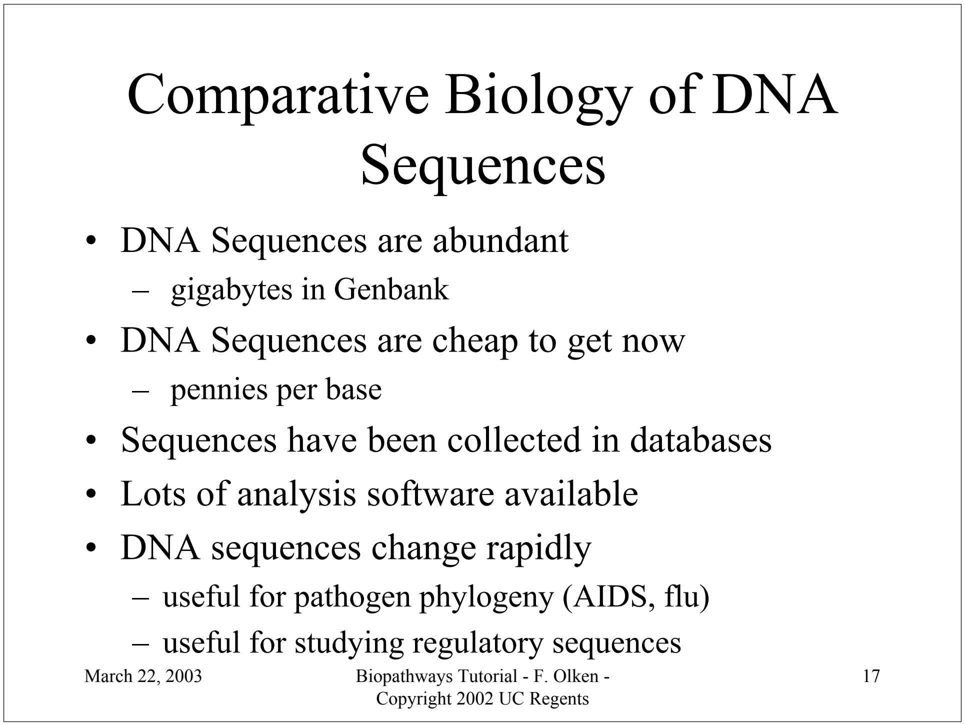
Uses of Biopathways DB

- Formal encoding of biological knowledge
- Drive simulations (add math models)
- Access to literature (links from reactions)
- Access to data (microarray data)
- Assist gene annotation
  - coregulation of genes (e.g., from microarray data) suggests participation in same pathways



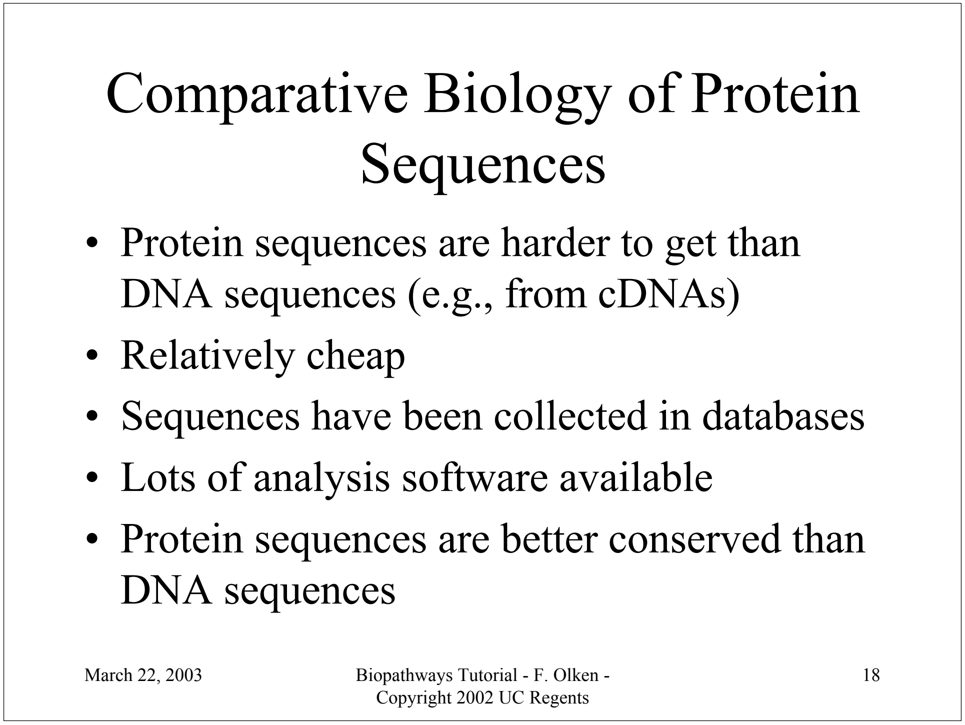
Comparative Biology on Biopathways

- Comparative analyses of DNA/RNA/protein sequences has proven very useful
- Biopathways datasets for many organisms are becoming available
- We can now begin to do comparative analyses of biopathways
- Formal encoding of biopathways needed to permit automated comparisons



Comparative Biology of DNA Sequences

- DNA Sequences are abundant
  – gigabytes in Genbank
- DNA Sequences are cheap to get now
  – pennies per base
- Sequences have been collected in databases
- Lots of analysis software available
- DNA sequences change rapidly
  – useful for pathogen phylogeny (AIDS, flu)
  – useful for studying regulatory sequences



Comparative Biology of Protein Sequences

- Protein sequences are harder to get than DNA sequences (e.g., from cDNAs)
- Relatively cheap
- Sequences have been collected in databases
- Lots of analysis software available
- Protein sequences are better conserved than DNA sequences

Comparative Biology of Protein Structures

- Protein structures are expensive to determine - thousands of dollars each (or more)
- Few are known - 20K in PDB, 10K distinct
- Some software to compare
- Expensive computations
- Collected in PDB database
- More highly conserved than protein sequences

Comparative Biology of Biopathways

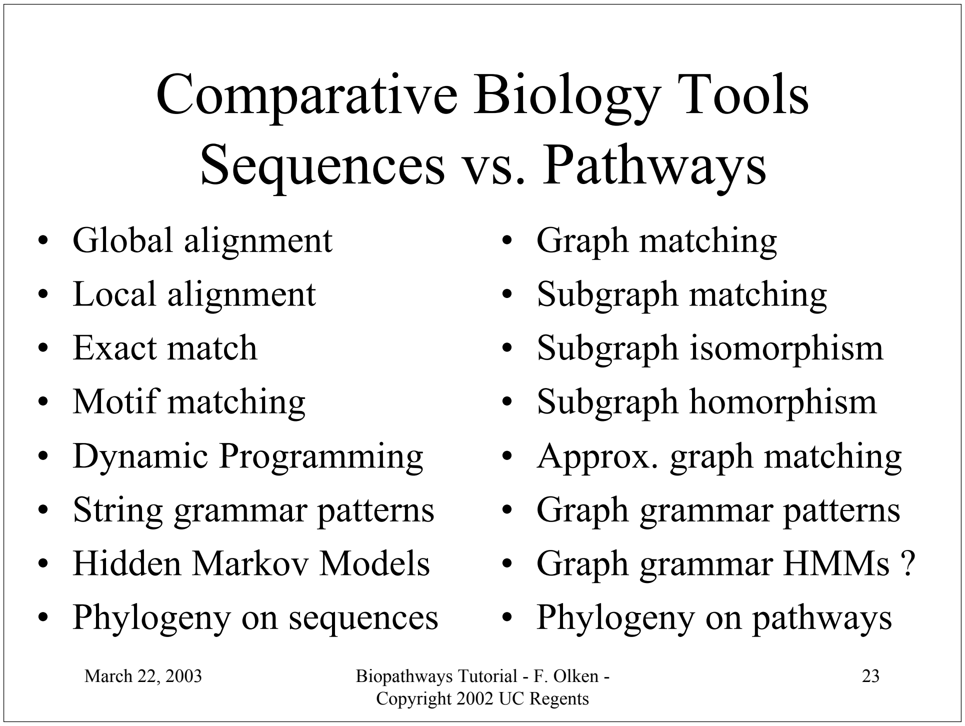
- **Small datasets**
  - **presently dozens of organisms (mostly microbes)**
  - **soon hundreds of organisms**
  - **complete ???**
- **Expensive to generate**
  - **often requires micro-array experiments, wet chemistry, gene knockouts, …**
- **Multiple databases**
- **Limited analysis software**

Comparative Analyses of Biopathways

- Biopathways are more strongly conserved than protein structures, protein sequences, or DNA sequences
- Hence, can be used for deep phylogeny

Requirements for pathway comparative biology

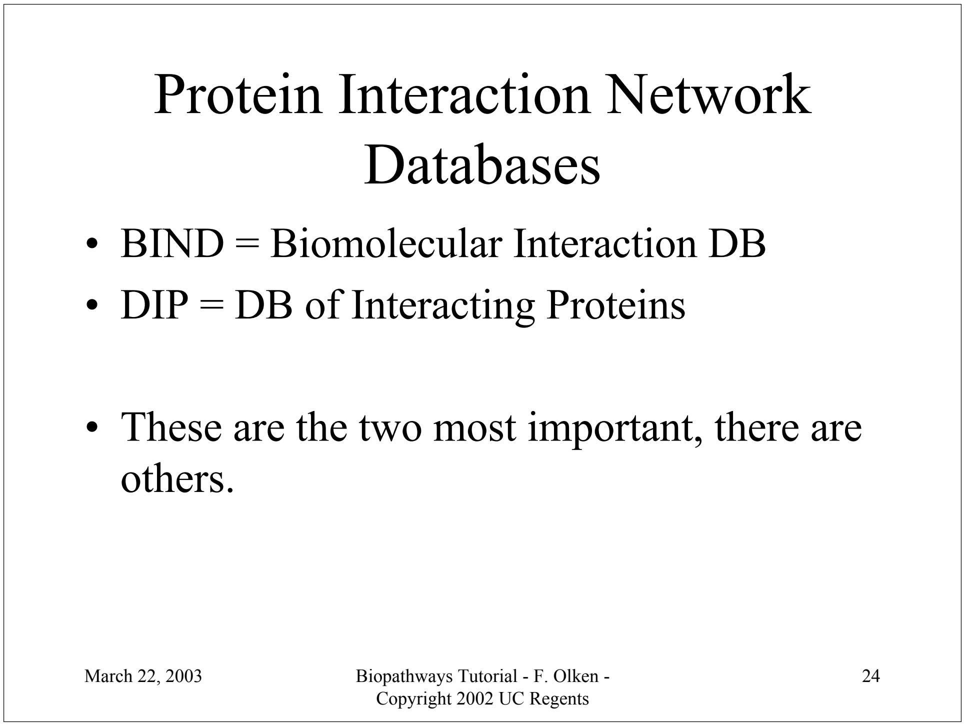
- Data (collections of biopathways)
- Data encoding (e.g., graphs)
- Algorithms for graph comparison
- Encodings for graph patterns
- Algorithms for graph pattern matching
- Algorithms for graph pattern discovery
- Tools for visualization of graph matchings



Comparative Biology Tools Sequences vs. Pathways

• Global alignment	• Graph matching
• Local alignment	• Subgraph matching
• Exact match	• Subgraph isomorphism
• Motif matching	• Subgraph homorphism
• Dynamic Programming	• Approx. graph matching
• String grammar patterns	• Graph grammar patterns
• Hidden Markov Models	• Graph grammar HMMs ?
• Phylogeny on sequences	• Phylogeny on pathways

March 22, 2003 Biopathways Tutorial - F. Olken - Copyright 2002 UC Regents 23

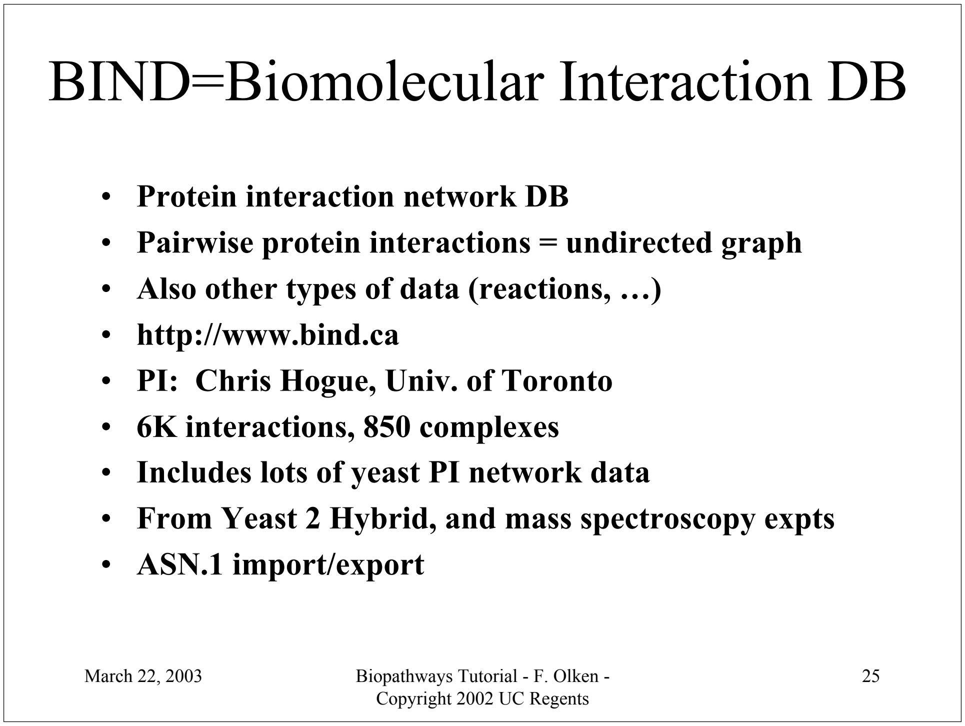


Protein Interaction Network Databases

- BIND = Biomolecular Interaction DB
- DIP = DB of Interacting Proteins

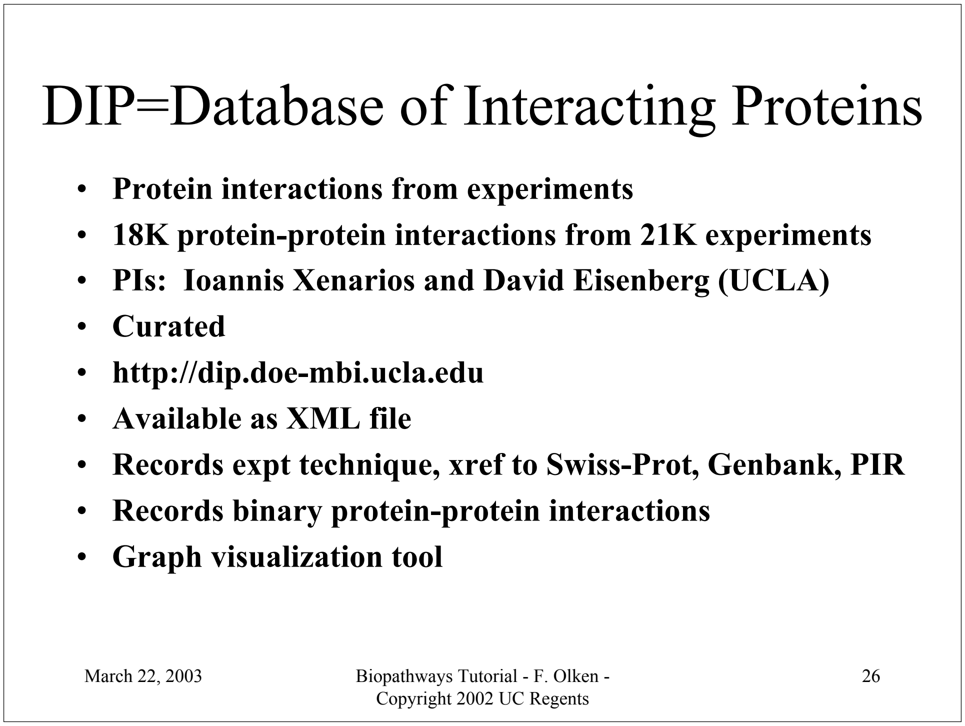
- These are the two most important, there are others.

March 22, 2003 Biopathways Tutorial - F. Olken - Copyright 2002 UC Regents 24



BIND=Biomolecular Interaction DB

- Protein interaction network DB
- Pairwise protein interactions = undirected graph
- Also other types of data (reactions, …)
- http://www.bind.ca
- PI: Chris Hogue, Univ. of Toronto
- 6K interactions, 850 complexes
- Includes lots of yeast PI network data
- From Yeast 2 Hybrid, and mass spectroscopy expts
- ASN.1 import/export



DIP=Database of Interacting Proteins

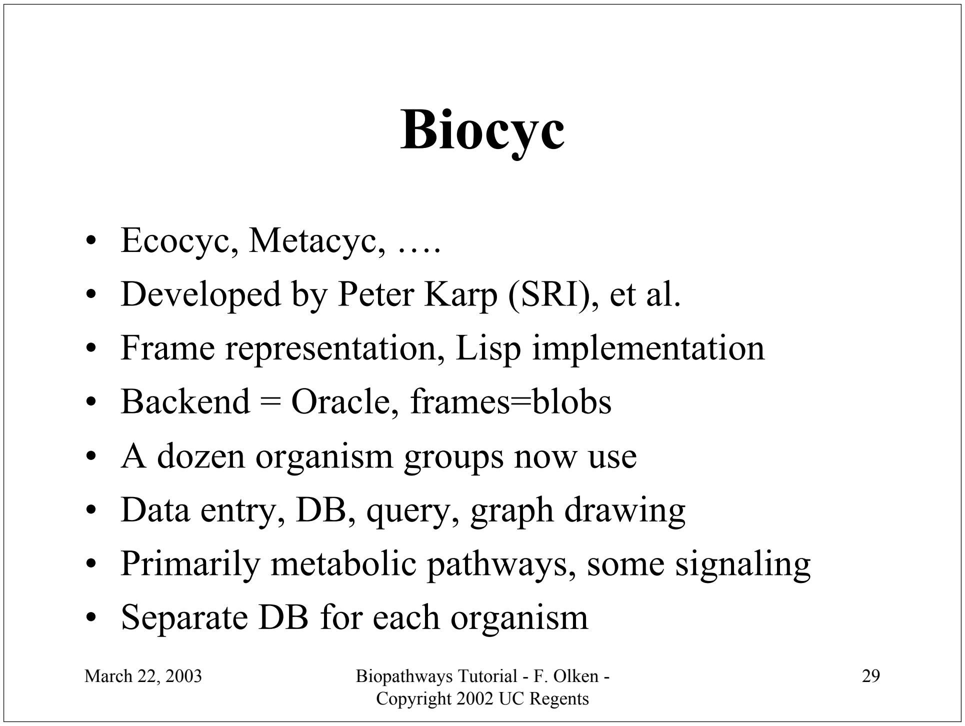
- Protein interactions from experiments
- 18K protein-protein interactions from 21K experiments
- PIs: Ioannis Xenarios and David Eisenberg (UCLA)
- Curated
- http://dip.doe-mbi.ucla.edu
- Available as XML file
- Records expt technique, xref to Swiss-Prot, Genbank, PIR
- Records binary protein-protein interactions
- Graph visualization tool

Survey Paper on Protein Interaction Databases

- Xenarios, I., and D. Eisenberg, “Protein Interaction Databases”, *Current Opinions in Biotechnology,* vol. 12, pp. 334-339

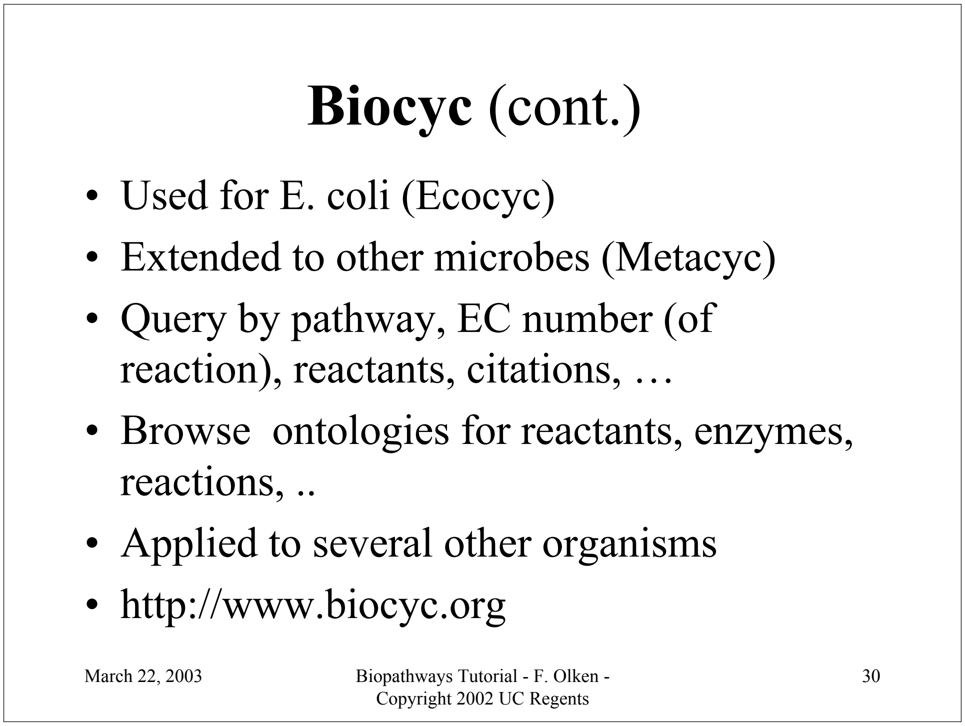
Biopathways Databases

- Biocyc: Ecocyc, …
- KEGG
- EMP, WIT, …
- Klotho
- aMAZE
- BGDM - Biopathways Graph Data Manager



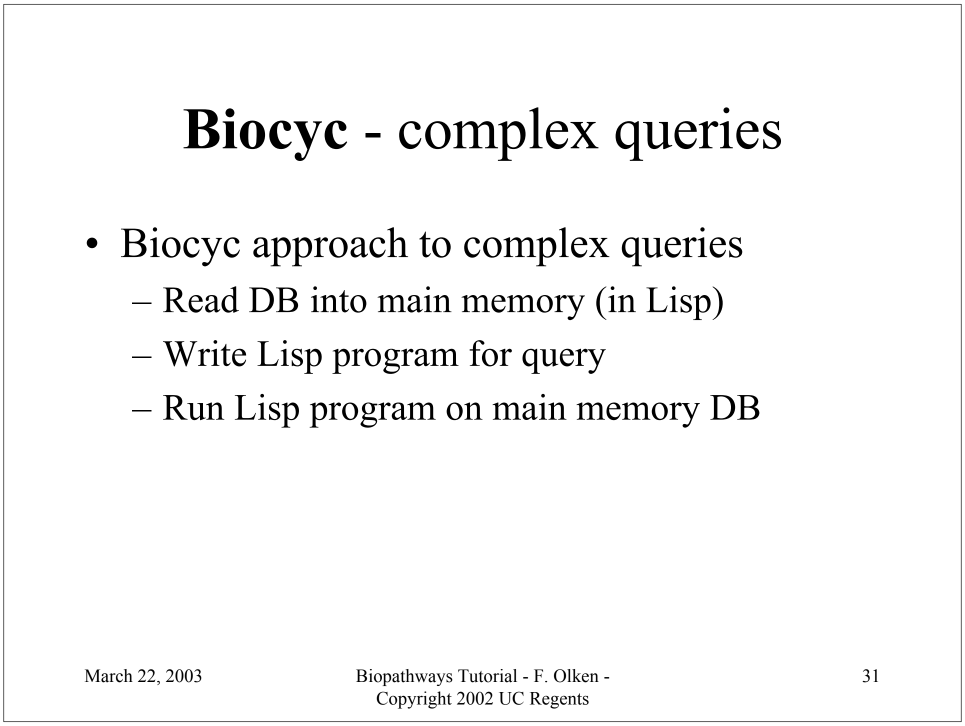
Biocyc

- Ecocyc, Metacyc, ….
- Developed by Peter Karp (SRI), et al.
- Frame representation, Lisp implementation
- Backend = Oracle, frames=blobs
- A dozen organism groups now use
- Data entry, DB, query, graph drawing
- Primarily metabolic pathways, some signaling
- Separate DB for each organism



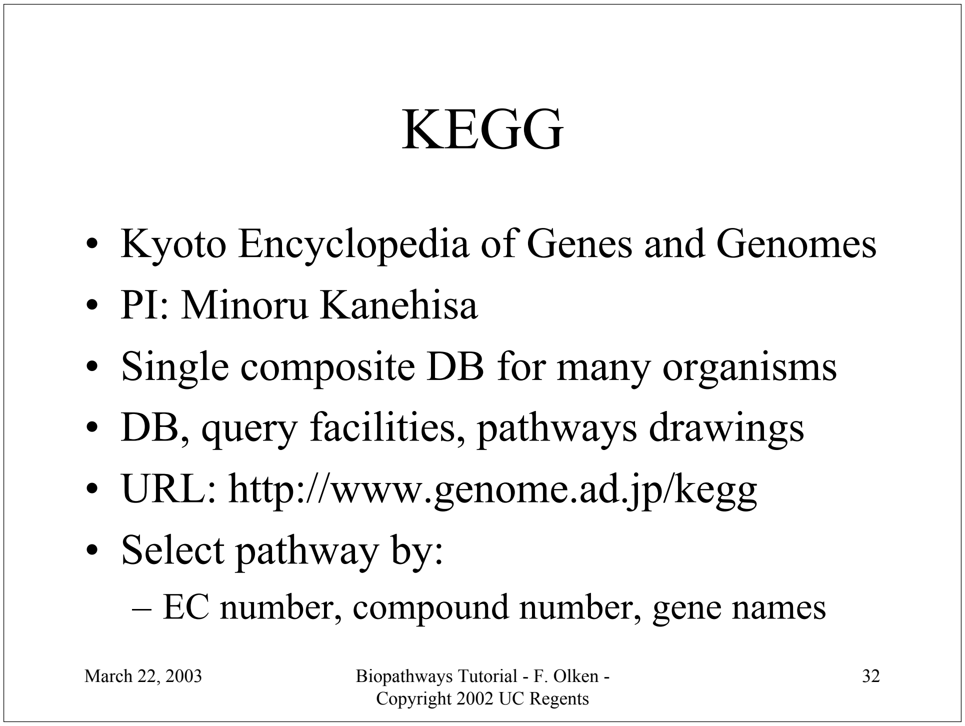
Biocyc (cont.)

- Used for E. coli (Ecocyc)
- Extended to other microbes (Metacyc)
- Query by pathway, EC number (of reaction), reactants, citations, …
- Browse ontologies for reactants, enzymes, reactions, ..
- Applied to several other organisms
- http://www.biocyc.org



Biocyc - complex queries

- Biocyc approach to complex queries
  - Read DB into main memory (in Lisp)
  - Write Lisp program for query
  - Run Lisp program on main memory DB



KEGG

- Kyoto Encyclopedia of Genes and Genomes
- PI: Minoru Kanehisa
- Single composite DB for many organisms
- DB, query facilities, pathways drawings
- URL: http://www.genome.ad.jp/kegg
- Select pathway by:
  - EC number, compound number, gene names

EMP, WIT, WIT2

- Eugeni Selkov, Eugeni Selkov, Jr., et al.
- Developed orginally in Russia
- Now at Argonne and Integrated Genomics
- Query pathway by substrate, enzyme, end product, ...
- Servers at IG, Argonne, …
- http://wit.mcs.anl.gov/WIT2

Klotho: Biochemical Compunds Declarative Database

- Developed by Toni Kazic (U. Missouri at Columbia)
- Chemical structure graphs of reactants
- Detailed modeling of chemical reaction mechanisms (cf. stoichiometry only in other DB)
- Written in Prolog
- Public server, DB, open source
- http://www.biocheminfo.org/klotho

aMAZE

- Metabolic + regulatory pathways database
- Developed by EBI in England
- URL: http://www.ebi.ac.uk/research/pfmp/
- Shoshana Wodak is PI
- Object Oriented DB
- Entity-association model
- Entities: metabolites, proteins, genes, …
- Associations: reactions, catalysis, transport

aMaze Queries

- Paths: find all paths from A to B
- Pattern search:
  - – branch points, feedback loops,
  - – pathways affected by a transcription factor
- Pattern discovery (?)
- High level abstraction (?)

Biopathways Graph Data Manager

- New project at LBNL
- Graph-based data manager
- Graph data model
- Graph queries
- Funded by DOE GTL and DARPA Biospice
- URL:
  - **http://www.lbl.gov/~olken/graphdm/graphdm.htm**

Biopathways Graph Data Model

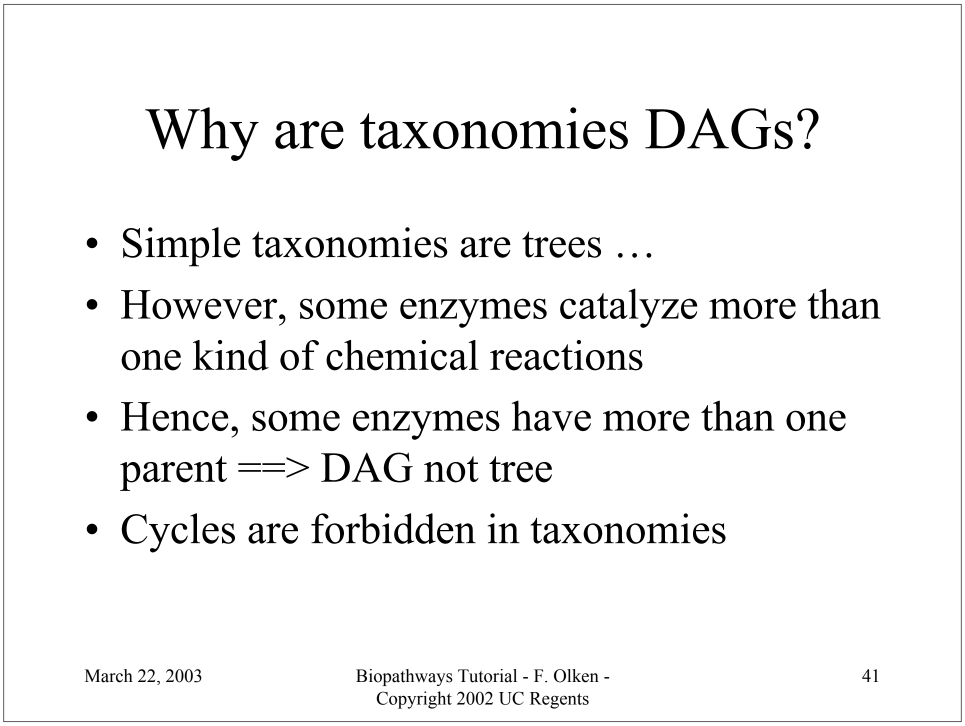
- Nodes:
  - Chemical entities: metabolites, enzymes, …
  - Bioprocesses: reactions, gene expression, …
- Edges (directed)
  - Indicate relationships
    - input, output, catalyze, inhibit, promote
    - is-a, part-of, element-of, ...

Taxonomies

- Of enzymes, reactants, organisms, …
- Simplest are hierarchies (trees)
  - each node has exactly one parent (except root)
  - like library classification systems
- Realistic taxonomies are often DAGs
  - directed acyclic graphs (no cycles)
  - nodes may have multiple parents
- Taxonomies specify partial orders

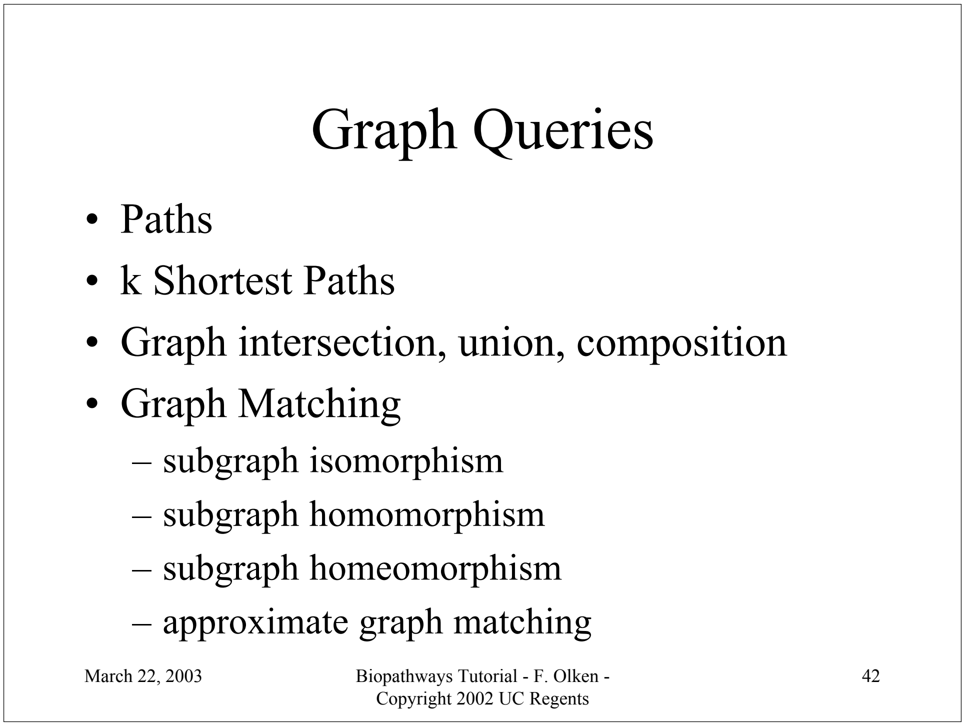
Role of Taxonomies

- Organism, enzyme, metabolite, reaction taxonomies
- Graphs are DAGs (directed acyclic graphs)
- Directed edges = is-a, instance-of
- Node labels are terms
- Terms are used to label nodes in query subgraphs
- Generic terms (upper levels of taxonomy) in query subgraphs must be expanded before performing subgraph matching.
- Example: find reaction containing a kinase enzyme



Why are taxonomies DAGs?

- Simple taxonomies are trees …
- However, some enzymes catalyze more than one kind of chemical reactions
- Hence, some enzymes have more than one parent ==> DAG not tree
- Cycles are forbidden in taxonomies



Graph Queries

- Paths
- k Shortest Paths
- Graph intersection, union, composition
- Graph Matching
  - – subgraph isomorphism
  - – subgraph homomorphism
  - – subgraph homeomorphism
  - – approximate graph matching

More Graph Queries

- Transitive Closure
- Least Common Ancestor
- Largest Common Subgraph

Paths

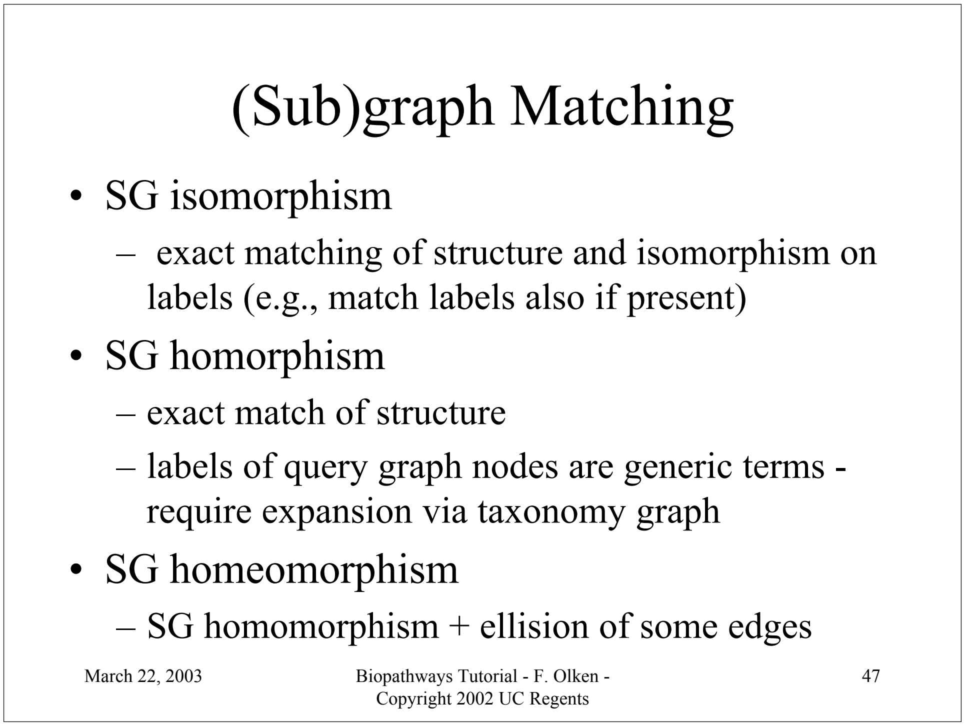
- Path = connected linear graph
- Example: A--->B--->C--->D
- Length(path) =
  - sum of “lengths of edges” along path
  - typically length of edge = 1

Path Queries

- Fixed length patterns
  - commonplace in Object oriented DB, XML DB
- Regular expressions on paths
  - matching labels on nodes (and edges)
  - recursive, arbitrary length paths
  - see work of Mendelzon, etc.
- Shortest path queries
  - k-shortest paths used a surrogate for most important paths in pathways DB
  - well known algorithms

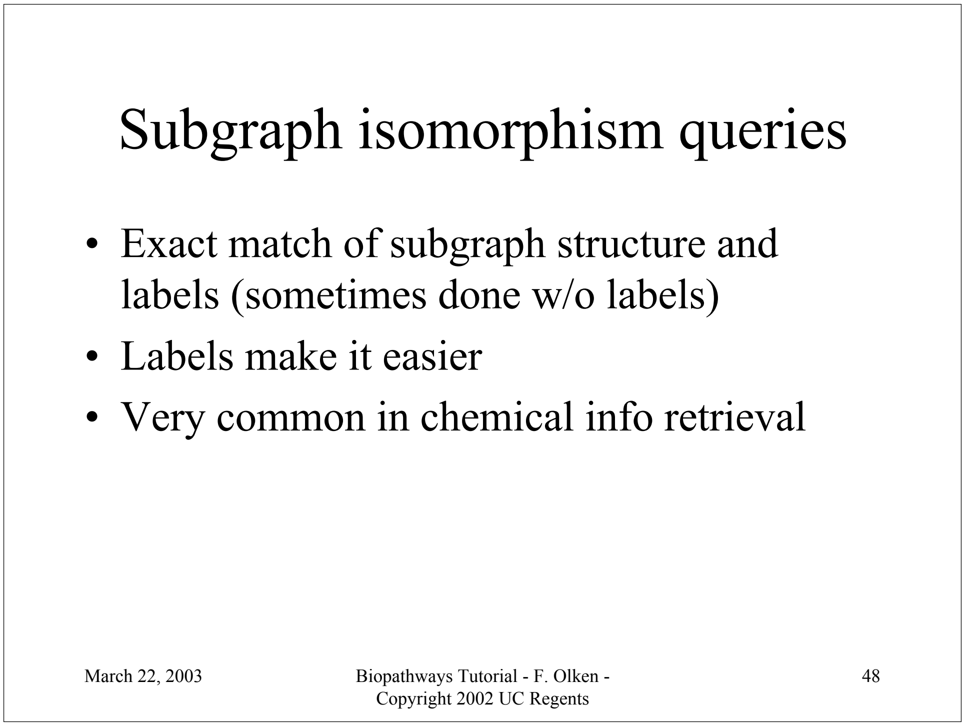
Neighborhood Queries

- Neighborhood of radius “r” of subgraph SG of graph G
- Subgraph H of G such that every node in H is within distance “r” of subgraph SG
- distance “r” = length of shortest path
  - edge lengths = 1
- Effect is to include portion of G which is near subgraph SG



(Sub)graph Matching

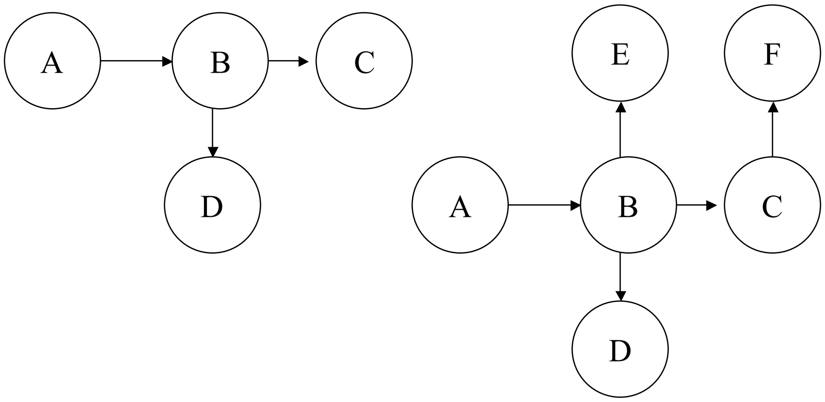
- SG isomorphism
  - exact matching of structure and isomorphism on labels (e.g., match labels also if present)
- SG homorphism
  - exact match of structure
  - labels of query graph nodes are generic terms - require expansion via taxonomy graph
- SG homeomorphism
  - SG homomorphism + ellision of some edges



Subgraph isomorphism queries

- Exact match of subgraph structure and labels (sometimes done w/o labels)
- Labels make it easier
- Very common in chemical info retrieval

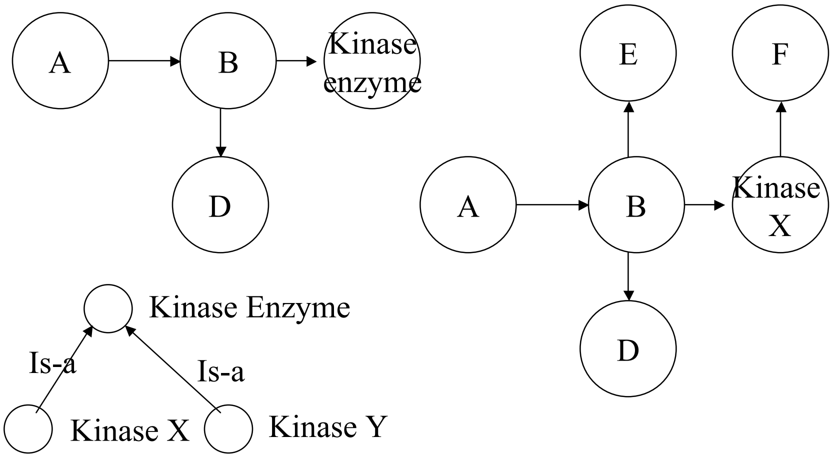
Example subgraph isomorphism



Subgraph homomorphism queries

- Exact match structure (edges)
- Labels on query subgraph are generic terms, e.g., kinase enzyme
- Matching labels are more specific terms subsumed by query node labels, e.g., particular kinase enzymes

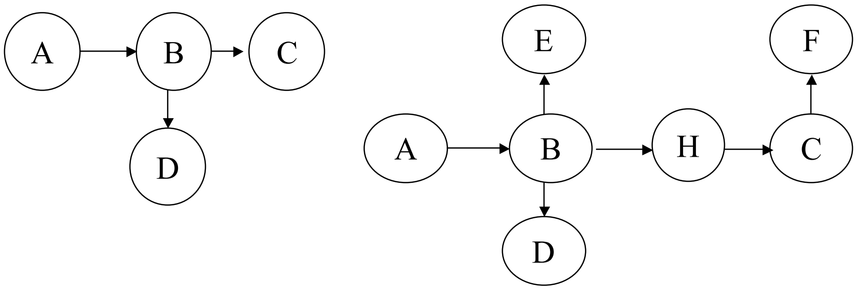
Example subgraph homomorphism



Subgraph homeomorphism

- Match a subgraph G in a graph H, by performing a subgraph isomorphism test against a “contraction of H”
- Contraction of H = contraction of some edge disjoint paths to single edges

Example subgraph homeomorphism



Approximate Graph Matching

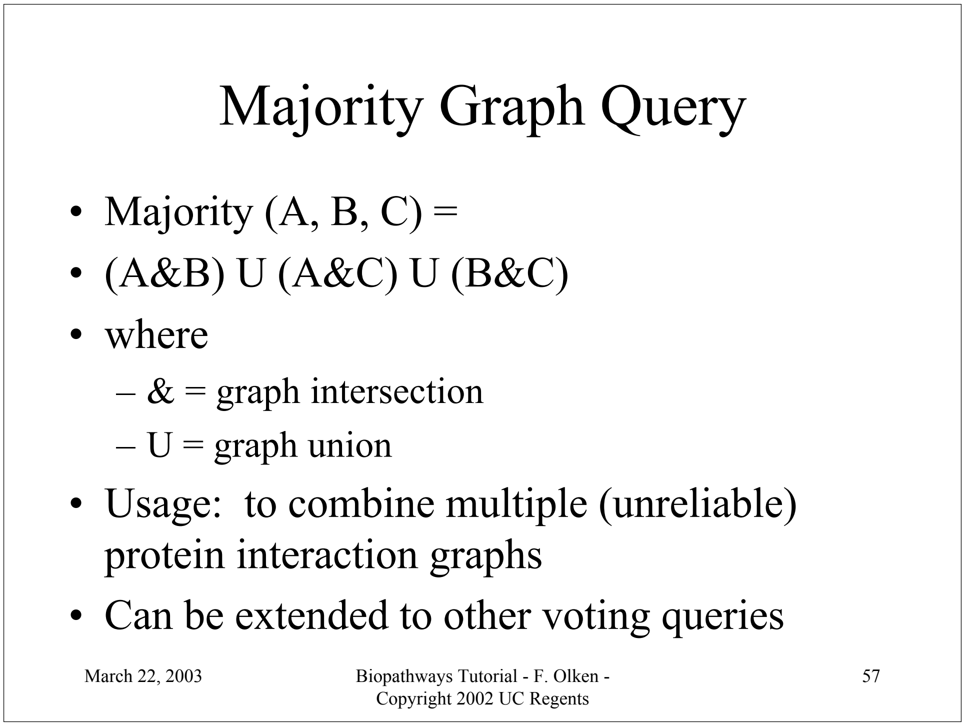
- See work by Dennis Shasha, et al.
- Akin to approximate string matching
- Allow: insertion, deletion, substitution of
  - nodes, edges, subgraphs
  - cost for each change
- Dynamic Programming used to find min. cost transformation from graph A to graph B

Node matching

- Needed for (sub)graph matching
- Node match via:
  - exact match of name
  - graph isomorphism of chemical structure graph associated with node
  - approx. string match of sequence (protein, DNA, …)
  - precomputed bipartite matching graph among nodes
  - various algorithmic definitions
  - match node labels + match surrounding context

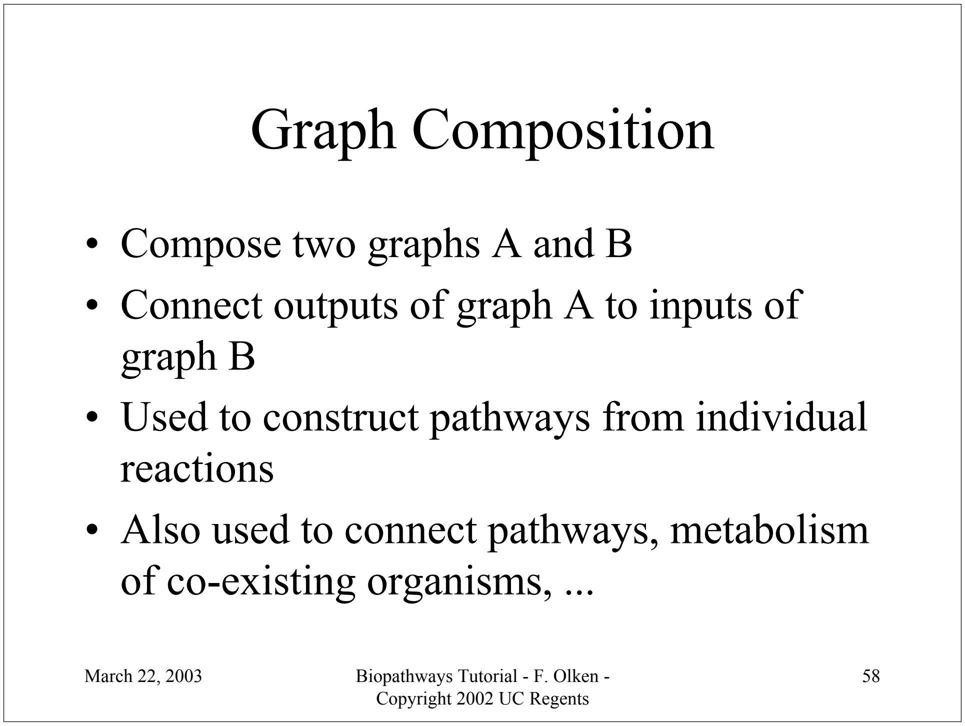
Boolean Graph Queries

- Graph intersection, union, difference
- Take intersection, union, difference, … of node sets, edge sets
- Note: graph intersection and union can be used to construct majority voting over 3 graphs
- Application: find the difference in metabolisms between two microbes



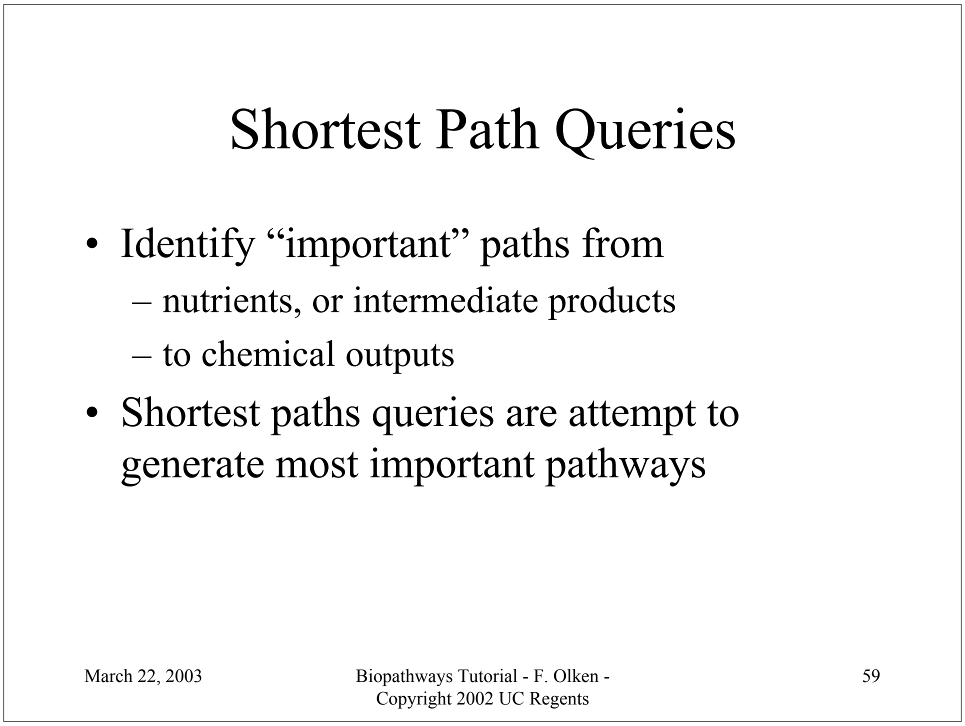
Majority Graph Query

- Majority (A, B, C) =
- (A&B) U (A&C) U (B&C)
- where
  - & = graph intersection
  - U = graph union
- Usage: to combine multiple (unreliable) protein interaction graphs
- Can be extended to other voting queries



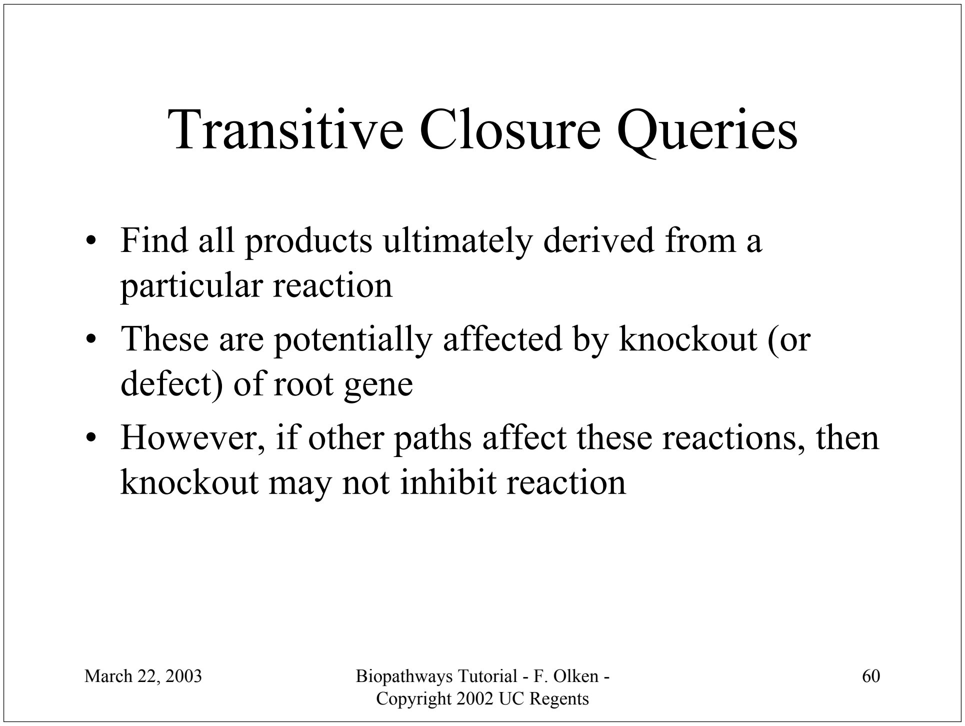
Graph Composition

- Compose two graphs A and B
- Connect outputs of graph A to inputs of graph B
- Used to construct pathways from individual reactions
- Also used to connect pathways, metabolism of co-existing organisms, ...



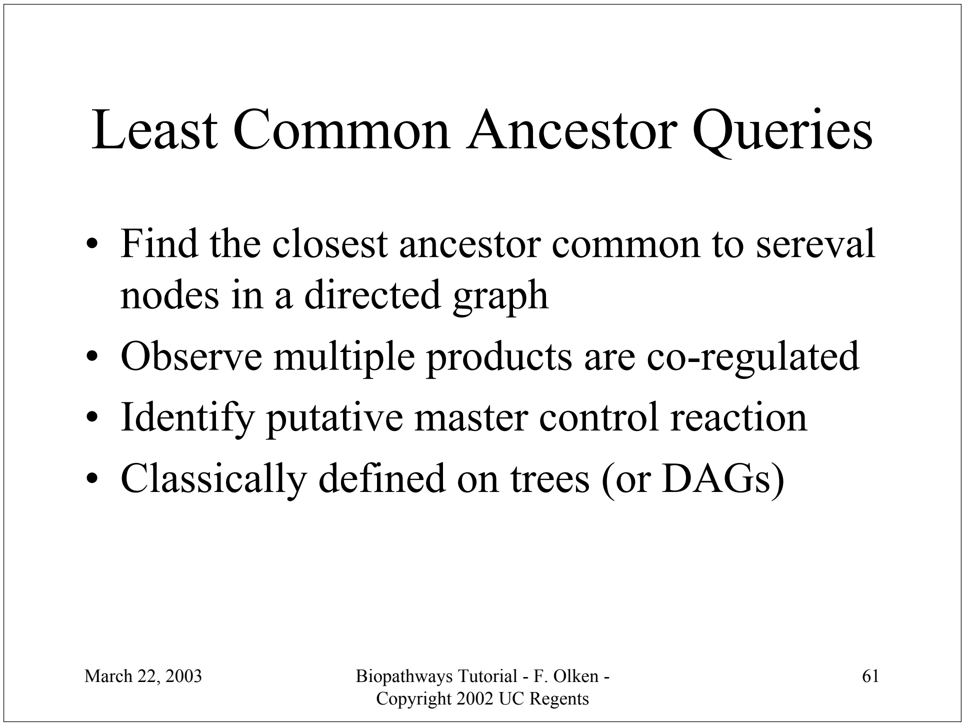
Shortest Path Queries

- Identify “important” paths from
  - – nutrients, or intermediate products
  - – to chemical outputs
- Shortest paths queries are attempt to generate most important pathways



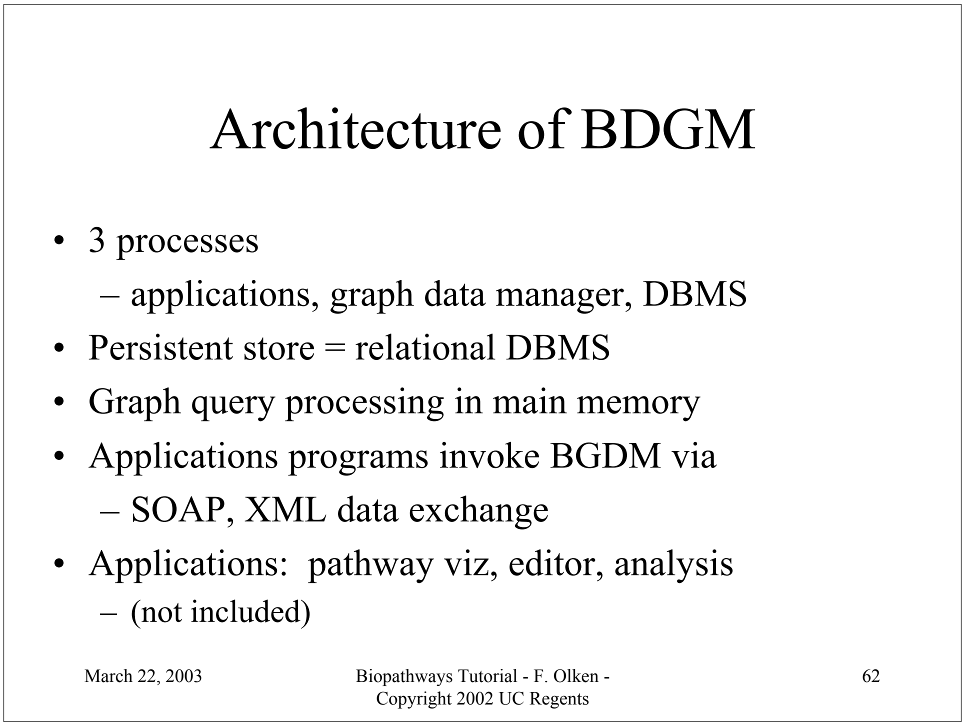
Transitive Closure Queries

- Find all products ultimately derived from a particular reaction
- These are potentially affected by knockout (or defect) of root gene
- However, if other paths affect these reactions, then knockout may not inhibit reaction



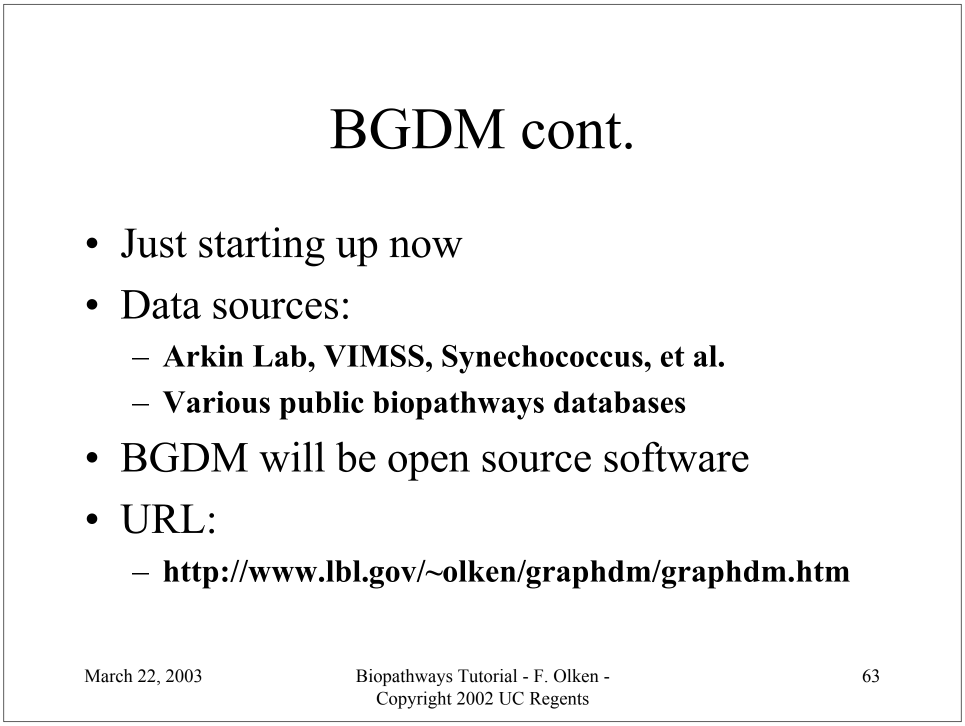
Least Common Ancestor Queries

- Find the closest ancestor common to sereval nodes in a directed graph
- Observe multiple products are co-regulated
- Identify putative master control reaction
- Classically defined on trees (or DAGs)



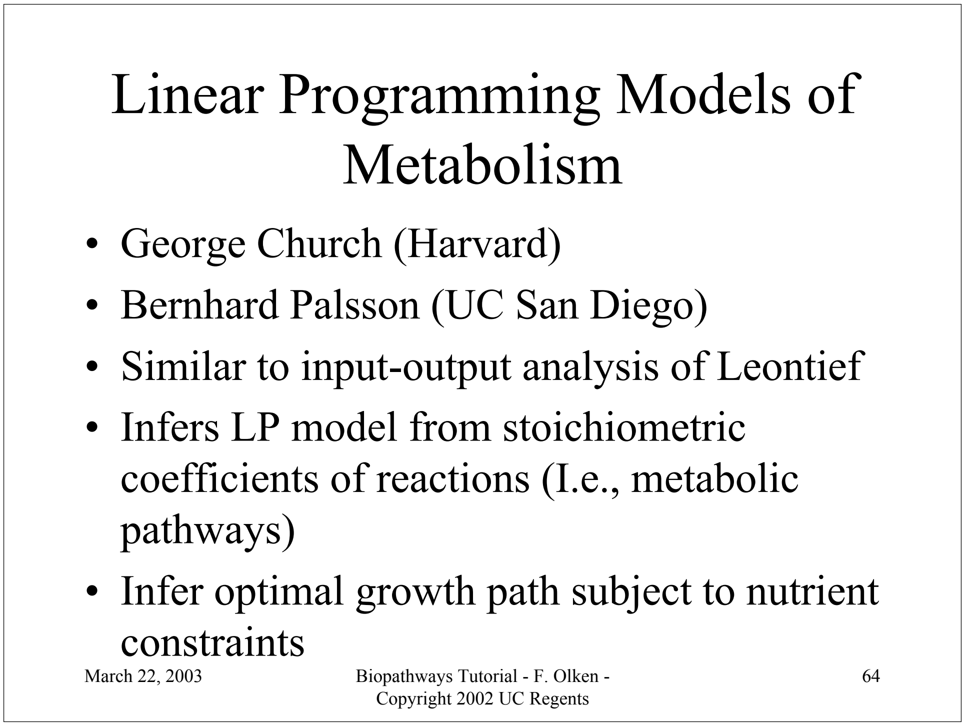
Architecture of BDGM

- 3 processes
  - applications, graph data manager, DBMS
- Persistent store = relational DBMS
- Graph query processing in main memory
- Applications programs invoke BGDM via
  - SOAP, XML data exchange
- Applications: pathway viz, editor, analysis
  - (not included)



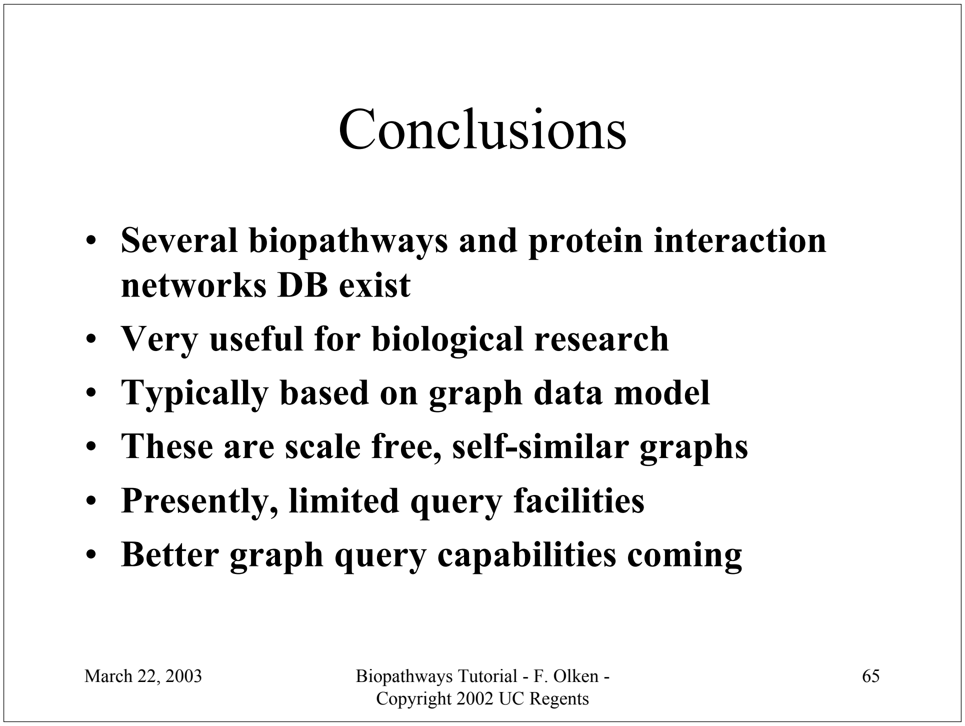
BGDM cont.

- Just starting up now
- Data sources:
  - **Arkin Lab, VIMSS, Synechococcus, et al.**
  - **Various public biopathways databases**
- BGDM will be open source software
- URL:
  - **http://www.lbl.gov/~olken/graphdm/graphdm.htm**



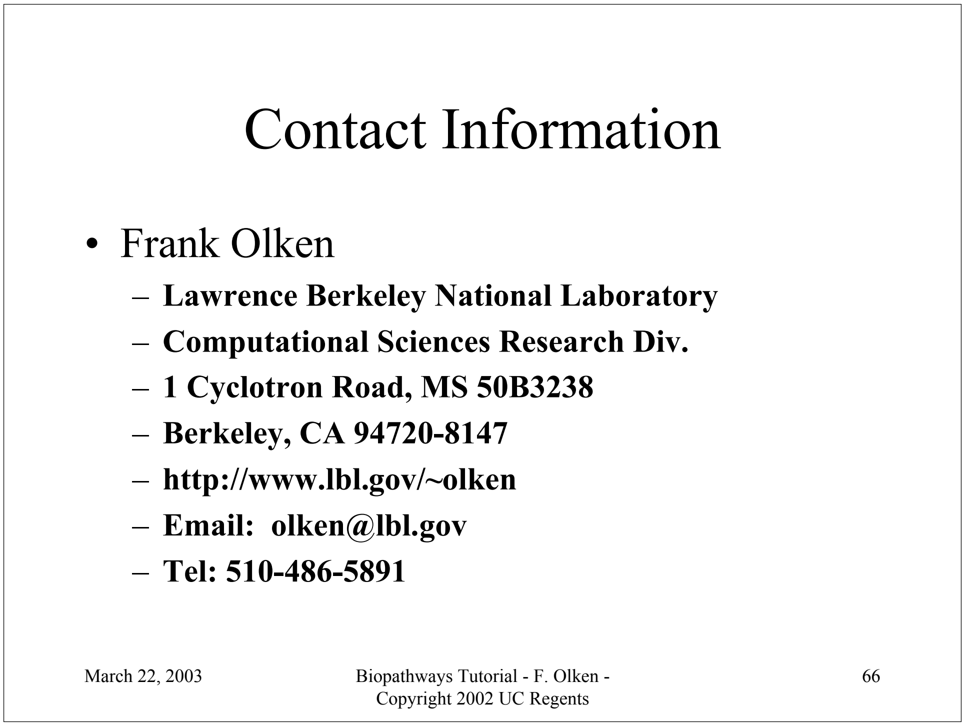
Linear Programming Models of Metabolism

- George Church (Harvard)
- Bernhard Palsson (UC San Diego)
- Similar to input-output analysis of Leontief
- Infers LP model from stoichiometric coefficients of reactions (I.e., metabolic pathways)
- Infer optimal growth path subject to nutrient constraints



Conclusions

- Several biopathways and protein interaction networks DB exist
- Very useful for biological research
- Typically based on graph data model
- These are scale free, self-similar graphs
- Presently, limited query facilities
- Better graph query capabilities coming



Contact Information

- Frank Olken
  - **Lawrence Berkeley National Laboratory**
  - **Computational Sciences Research Div.**
  - **1 Cyclotron Road, MS 50B3238**
  - **Berkeley, CA 94720-8147**
  - **http://www.lbl.gov/~olken**
  - **Email: olken@lbl.gov**
  - **Tel: 510-486-5891**

Funding Acknowledgements

- **VIMSS (LBNL Genome to Life Project)**
  - **funded by Dept. of Energy, OBER**
  - **A. Arkin is PI**
- **Synechococcus (Sandia GTL Project)**
  - **funded by DOE OBER, OASCR**
  - **G. Heffelfinger PI (A. Shoshani LBNL PI)**
- **Berkeley BIOSPICE**
  - **funded by DARPA**
  - **A. Arkin is PI**

Personal Acknowledgements

- **Manfred Zorn (LBNL) - assist with figures**
- **Jean Faulon (Sandia) - Network characterization**
- **Sylvia Spengler (NSF) - encouraged work on graph databases**
- **Adam Arkin (LBNL) - supports graph DB work**
- **M. Fernandez, J. Simeone, P. Wadler - Xquery, functional programming approaches to query languages**